组会

21.03.20 许晓丹

AAAI 2020

Improving Question Generation with Sentence-level Semantic Matching and Answer Position Inferring

Xiyao Ma[1], **Qile Zhu**[1], **Yanlin Zhou**[1], **Xiaolin Li**[2], **Dapeng Wu**[1]
[1]NSF Center for Big Learning, University of Florida
[2]AI Institute, Tongdun Technology
maxiy@ufl.edu, valder@ufl.edu, zhou.y@ufl.edu, xiaolin.li@tongdun.net, dpwu@ufl.edu

Motivation

1. 通过分析以往的QG SOTA模型发现，现存模型的生成主要有两个问题：
   1）错误的疑问词或关键词；
   2）copy机制会复制与答案语义无关的词。
2. 解码器忽略了问题的全局语义，copy机制未能很好地利用答案的位置特征。

以Multi-Task Learning的方式，在正常的seq2seq外，添加了两个模块：

句子级的语义匹配模块 + 答案位置推断模块

Dataset

基于SQuAD 和 MS MARCO；

句子级：sentence-answer-question

Model

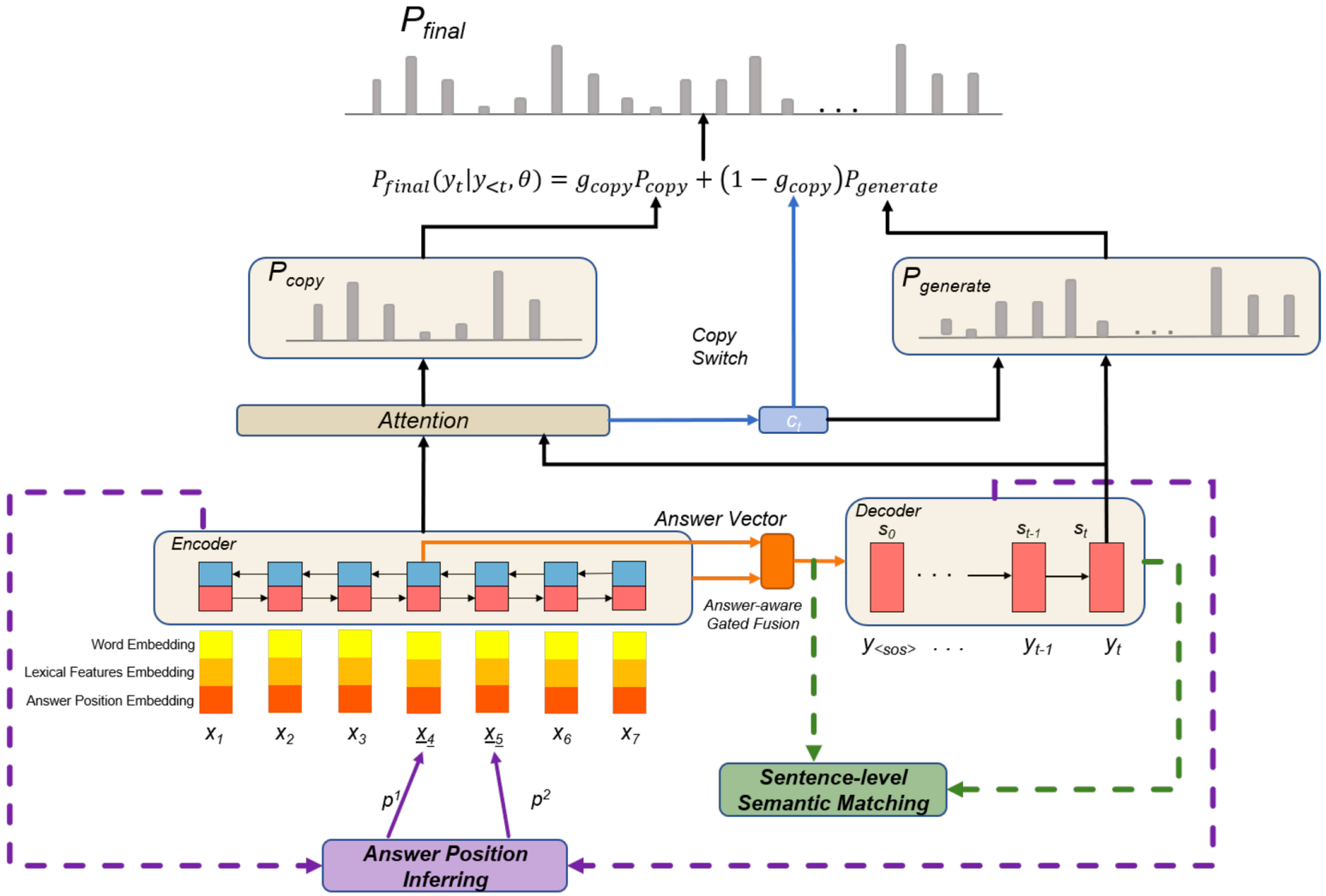


Figure 1: Diagram for neural question generation model with sentence-level semantic matching, answer position inferring, and gated fusion.

Vanilla Model Structure

Encoder: Bi-LSTM

Decoder: LSTM with attention mechanism

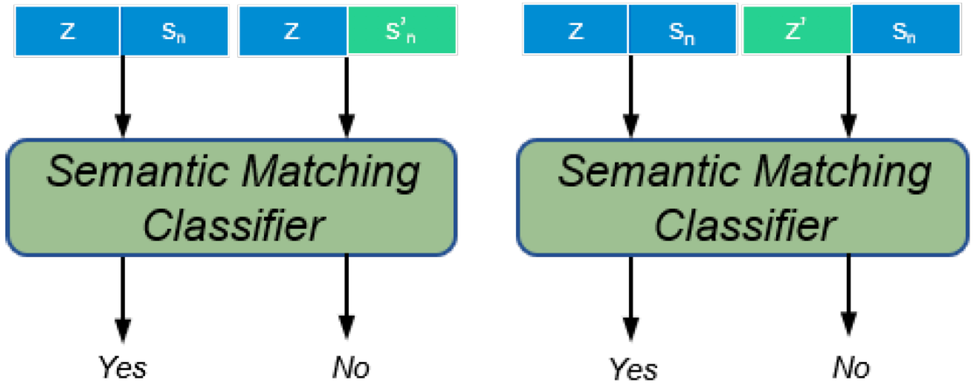
Answer-aware Gated Fusion:

$$g_m = \sigma(W_m^T * [h_m, h_a] + b_m),$$
$$g_a = \sigma(W_a^T * [h_m, h_a] + b_a),$$
$$z = g_m \cdot h_m + g_a \cdot h_a$$

answer-aware sentence vector

Copy mechanism

Sentence-Level Semantic Matching



<sentence, answer1, question1> <sentence, answer2, question2>

z: answer-aware 句子向量；

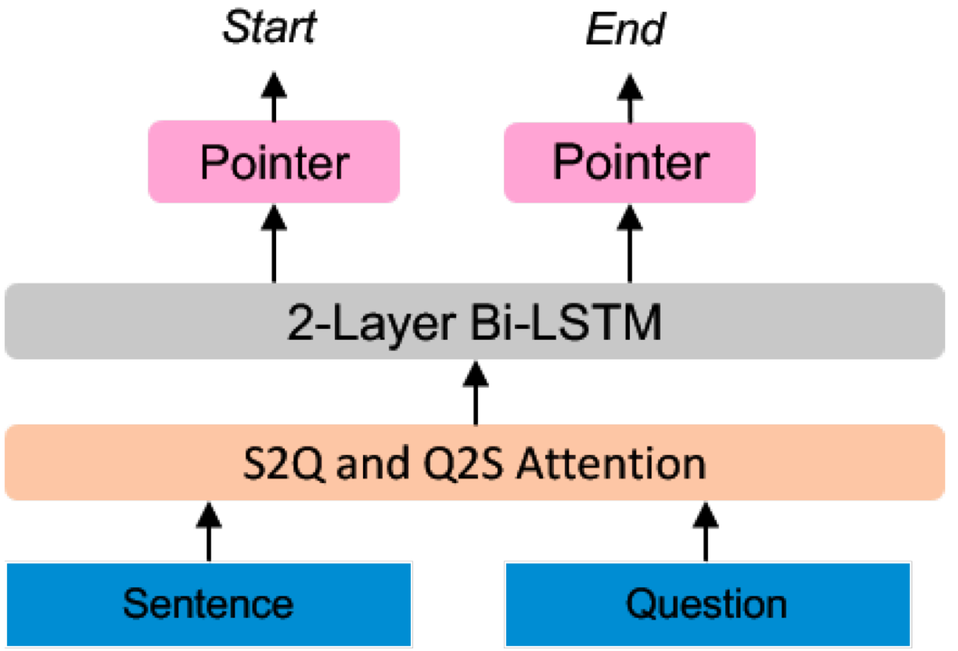
sn: 把decoder看作是question的encoder，那么sn可被视作question vector.

$$p_{sm} = Softmax(W_c[z^i, s_n^i] + b_c)$$

Answer Position Inferring

引入双向注意流推断答案位置：

$$\tilde{H} = attn(H, S),$$
$$\tilde{S} = attn(S, H)$$
$$M_1 = LSTM(f(H, \tilde{H}, \tilde{S})),$$
$$M_2 = LSTM(M_1),$$
$$p^1 = \text{Softmax}\left(W_{(p^1)}^{\top}[\tilde{H}, M_1]\right),$$
$$p^2 = \text{Softmax}\left(W_{(p^2)}^{\top}[\tilde{H}, M_2]\right)$$



Experiments

Table 3: Comparison of models performances in terms of the main met

	SQuAD					
Models	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
NQG++ (Zhou et al. 2017)	42.13	25.98	18.24	13.29	17.59	40.75
Pointer-generator (See, Liu, and Manning 2017)	42.43	26.75	18.99	14.33	18.77	43.19
Answer-focused (Sun et al. 2018)	43.02	28.14	20.51	15.64	-	-
Gated Self-attention (Zhao et al. 2018)	44.51	29.07	21.06	15.82	19.67	44.24
Model with Sentence-level Semantic Matching	43.67	28.53	20.59	15.66	19.23	43.86
Model with Answer Position Inferring	43.88	28.55	28.87	15.77	19.55	43.98
Combined Model	**44.71**	**29.89**	**21.77**	**16.32**	**20.84**	**44.79**

Table 4: Machine Comprehension Performance in terms of Exact Match (EM) and F1on SQUAD dataset

Questions	EM (%)	F1 (%)
Reference Questions	49.68	65.97
NQG++ (Zhou et al. 2017)	35.26	50.88
Pointer-generator (See, Liu, and Manning 2017)	38.89	54.06
Our model	**42.70**	**57.68**

Experiments

Table 5: Question words and keywords generation performance by different models on SQuAD dataset

Models	# right question words	# right keywords
NQG++ (Zhou et al. 2017)	134	143
Pointer-generator (See, Liu, and Manning 2017)	140	148
Model with Sentence-level Semantic Matching	**150**	**156**

Table 6: Copy mechanism performance by different models

Models	Precision	Recall
NQG++ (Zhou et al. 2017)	46.28%	32.13%
Pointer-generator (See, Liu, and Manning 2017)	47.21%	38.38%
Model with Answer Position Inferring	**48.35%**	**40.27%**

Experiments

Table 7: Performance Improvement on existing models on SQuAD dataset

Models	BLEU-4
NQG++ (Zhou et al. 2017)	13.29
NQG++ (Zhou et al. 2017) + our work	**14.97**
Pointer-generator model (See, Liu, and Manning 2017)	14.33
Pointer-generator model (See, Liu, and Manning 2017) + our work	**16.32**

NAACL 2021

Are NLP Models really able to Solve Simple Math Word Problems?

Arkil Patel **Satwik Bhattamishra** **Navin Goyal**
Microsoft Research India
{t-arkpat,t-satbh,navingo}@microsoft.com

Motivation

对于现有数据集，

1. MWP solvers不需要看到question就能求解相当一部分MWP；
2. 把MWP看做bag-of-words，模型也可以取得较高的准确率。

现存在的MWP solvers大部分依赖浅层的模式，

从而在benchmark 的数据集上得到很好的效果

Experiments——Mask Question

Model	MAWPS	ASDiv-A
Seq2Seq (S)	79.7	55.5
Seq2Seq (R)	86.7	76.9
GTS (S) (Xie and Sun, 2019)	82.6	71.4
GTS (R)	88.5	81.2
Graph2Tree (S) (Zhang et al., 2020)	83.7	77.4
Graph2Tree (R)	**88.7**	**82.2**
Majority Template Baseline	17.7	21.2

Table 2: 5-fold Cross Validation Accuracies (↑) of baseline models on datasets. (R) means that the model is provided with RoBERTa pretrained embeddings while (S) means that the model is trained from scratch.

Model	MAWPS	ASDiv-A
Seq2Seq	77.4	58.7
GTS	76.2	60.7
Graph2Tree	**77.7**	**64.4**

Table 3: 5-fold Cross Validation Accuracies (↑) of baseline models on Question-removed datasets

现有模型只是依赖于Body中简单的启发式模式来预测数学表达式

Experiments——Mask Question

	MAWPS		ASDiv-A	
Model	*Easy*	*Hard*	*Easy*	*Hard*
Seq2Seq	86.8	86.7	91.3	56.1
GTS	92.6	71.7	91.6	65.3
Graph2Tree	93.4	71.0	92.8	63.3

Table 4: Results of baseline models on the *Easy* and *Hard* test sets.

现有模型只是依赖于Body中简单的启发式模式

来预测数学表达式

Experiments——Remove Word-Order Information

将Seq2Seq模型中的LSTM Encoder用FFN代替 (non-contextual)

Model	MAWPS	ASDiv-A
FFN + LSTM Decoder (S)	75.1	46.3
FFN + LSTM Decoder (R)	**77.9**	**51.2**

Table 5: 5-fold Cross Validation Accuracies (↑) of the constrained model on the datasets. (R) denotes that the model is provided with non-contextual RoBERTa pre-trained embeddings while (S) denotes that the model is trained from scratch.

Experiments——Remove Word-Order Information

应用attention机制，观察解码时究竟attend到了什么

Input Problem	Predicted Equation	Answer
John delivered 3 letters at every house. If he delivered for 8 houses, how many letters did John deliver?	3 * 8	24 ✓
John delivered 3 letters at every house. He delivered 24 letters in all. How many houses did John visit to deliver letters?	3 * 24	72 ✗
Sam made 8 dollars mowing lawns over the Summer. He charged 2 bucks for each lawn. How many lawns did he mow?	8 / 2	4 ✓
Sam mowed 4 lawns over the Summer. If he charged 2 bucks for each lawn, how much did he earn?	4 / 2	2 ✗
10 apples were in the box. 6 are red and the rest are green. how many green apples are in the box?	10 - 6	4 ✓
10 apples were in the box. Each apple is either red or green. 6 apples are red. how many green apples are in the box?	10 / 6	1.67 ✗

Table 6: Attention paid to specific words by the constrained model.

只需要将MWP中special words的出现 映射到其对应的数学表达式

New Dataset——SVAMP

1. 现在benchmark数据集上的良好表现是misleading的；
2. 目标是建立一个简单的MWP的数据集，所有致力于解决MWP的系统

都应该在这个数据集上得到良好效果。

从ASDiv-A中选取seed sample，然后根据以下三类进行扩展：

1. Question Sensitivity
2. Reasoning Ability
3. Structural Invariance

数学表达式不超过2个操作符

Experiments

	Seq2Seq		GTS		Graph2Tree	
	S	*R*	*S*	*R*	*S*	*R*
Full Set	24.2	40.3	30.8	41.0	36.5	**43.8**
One-Op	25.4	42.6	31.7	44.6	42.9	**51.9**
Two-Op	20.3	**33.1**	27.9	29.7	16.1	17.8
ADD	28.5	**41.9**	35.8	36.3	24.9	36.8
SUB	22.3	35.1	26.7	36.9	41.3	**41.3**
MUL	17.9	**38.7**	29.2	**38.7**	27.4	35.8
DIV	29.3	56.3	39.5	61.1	40.7	**65.3**

Table 10: Results of models on the SVAMP challenge set. *S* indicates that the model is trained from scratch. *R* indicates that the model was trained with RoBERTa embeddings. The first row shows the results for the full dataset. The next two rows show the results for subsets of SVAMP composed of examples that have equations with one operator and two operators respectively. The last four rows show the results for subsets of SVAMP composed of examples of type Addition, Subtraction, Multiplication and Division respectively.

Model	SVAMP w/o ques	ASDiv-A w/o ques
Seq2Seq	29.2	58.7
GTS	28.6	60.7
Graph2Tree	**30.8**	**64.4**

Table 11: Accuracies (↑) of models on SVAMP without questions. The 5-fold CV accuracy scores for ASDiv-A without questions are restated for easier comparison.

Experiments

Model	SVAMP
FFN + LSTM Decoder (S)	17.5
FFN + LSTM Decoder (R)	**18.3**
Majority Template Baseline	11.7

Table 12: Accuracies (↑) of the constrained model on SVAMP. (R) denotes that the model is provided with non-contextual RoBERTa pretrained embeddings while (S) denotes that the model is trained from scratch. The Majority Template Baseline score is the accuracy on the dataset when the model always predicts the equation template with the highest frequency.

SVAMP不太容易被Body中简单的模式所解决，而且需要上下文信息

Experiments

Dataset	2 nums	3 nums	4 nums
ASDiv-A	93.3	59.0	47.5
SVAMP	78.3	25.4	25.4

Table 15: Accuracy break-up according to the number of numbers in the input problem. **2 nums** refers to the subset of problems which have only 2 numbers in the problem text. Similarly, **3 nums** and **4 nums** are subsets that contain 3 and 4 different numbers in the problem text respectively.

现有的模型没办法将number和他们的上下文联系到一起

Same Object, Different Structure	**Original:** Allan brought two balloons and Jake brought four balloons to the park. How many balloons did Allan and Jake have in the park? **Variation:** Allan brought two balloons and Jake brought four balloons to the park. How many more balloons did Jake have than Allan in the park?
Change Information	**Original:** Jack had 142 pencils. Jack gave 31 pencils to Dorothy. How many pencils does Jack have now? **Variation:** Dorothy had 142 pencils. Jack gave 31 pencils to Dorothy. How many pencils does Dorothy have now?
Change order of phrases	**Original:** Matthew had 27 crackers. If Matthew gave equal numbers of crackers to his 9 friends, how many crackers did each person eat? **Variation:** Matthew gave equal numbers of crackers to his 9 friends. If Matthew had a total of 27 crackers initially, how many crackers did each person eat?